PATH ANALYSIS WITH LOGISTIC REGRESSION MODELS: EFFECT ANALYSIS OF FULLY RECURSIVE CAUSAL SYSTEMS OF CATEGORICAL VARIABLES

Nobuoki Eshima*, Minoru Tabata** and Geng Zhi***

This paper discusses path analysis of categorical variables with logistic regression models. The total, direct and indirect effects in fully recursive causal systems are considered by using model parameters. These effects can be explained in terms of log odds ratios, uncertainty differences, and an inner product of explanatory variables and a response variable. A study on food choice of alligators as a numerical example is reanalysed to illustrate the present approach.

Key words and phrases: Categorical variable; Direct effect; Indirect effect; Log odds ratio; Logistic regression model; Recursive causal system; Total effect.

1. Introduction

Path analysis is usually performed for continuous variables by using linear regression equations (see Asher, 1976), and the basic idea is applied to the analysis of causal systems of continuous variables, LISREL model (Jöreskog and Sörbom, 1989). In comparison with path analysis of continuous variables, that of categorical variables is complex, because the causal system under consideration cannot be described by linear regression equations. Goodman (1973a, b, 1974) considered path analysis of binary variables by using logistic regression (logit) models (Cox, 1970), and discussed the effects by logit parameters. Hagenaars (1998) made a general discussion of path analysis of recursive causal systems of categorical variables by using the directed loglinear model approach, which is a combination of Goodman’s approach and graphical modeling. Although the approach is an analogy to LISREL approach, the discussion of the direct, indirect and total effects was not made. Eshima and Tabata (1999) discussed an effect analysis of recursive systems of categorical variables, and the effects of factors were defined by baseline log odds ratios. In path analysis, it is always a subject of discussion how the total, direct and indirect effects are defined (Fienberg, 1990; Hagenaars, 1998).

In this paper, path analysis is discussed in recursive causal systems shown in Fig. 1. Single head arrows indicate the direct effects. In con-
Continuous variables, the following model is assumed:

\[ X_k = \sum_{i=1}^{k-1} \beta_i X_i + \varepsilon_k \quad (k = 2, 3, \ldots, K), \]

where the variables \(X_i\) are standardized, and \(\varepsilon_k\) is the error term independent of explanatory variables \(X_i\). In this case, the direct effect of \(X_i\) on \(X_k\) is defined by

\[ e_d(X_i \rightarrow X_k) = \beta_i \]

and the total effect of \(X_i\) on \(X_k\) is defined by

\[ e_T(X_i \rightarrow X_k) = \beta_i + \sum_{j=i+1}^{k-1} \beta_j \rho_{ij}, \]

where \(\rho_{ij}\) are the correlation coefficients between \(X_i\) and \(X_j\). In this case, the interpretation of the effects is easy, and the indirect effect of \(X_i\) on \(X_k\) is defined by

\[ e_{ind}(X_i \rightarrow X_k) = \sum_{j=i+1}^{k-1} \beta_j \rho_{ij}. \]

In the case of categorical variables, let \(p(x_1, x_2, \ldots, x_k)\) be the joint probability of the random vector \((X_1, X_2, \ldots, X_k) = (x_1, x_2, \ldots, x_k)\), and \(p(x_i | x_1, x_2, \ldots, x_{i-1})(i = 2, 3, \ldots, k)\) be the conditional probabilities of \(X_i = x_i\) given \((X_1, X_2, \ldots, X_{i-1}) = (x_1, x_2, \ldots, x_{i-1})\). Then, we have the following decomposition of \(p(x_1, x_2, \ldots, x_k)\):

\[ p(x_1, x_2, \ldots, x_k) = p(x_1)p(x_2 | x_1) \cdots p(x_k | x_1, x_2, \ldots, x_{k-1}). \]

In this paper, conditional probabilities \(p(x_i | x_1, x_2, \ldots, x_{i-1})\) are assumed to be logistic regression models. The logistic regression model is the most frequently used regression model for categorical variables. The logistic regression models have been investigated by many authors, e.g. Anderson (1984), McCullagh (1980), Greenland (1994), Lian et al. (1992), Whittemore (1995) and others. Usually, the effects of explanatory variables on a response variable are assessed by using model parameters, however it is important in many applications to distinguish a direct effect from an
indirect effect of an explanatory variable on a response one. Parameters of logistic regression models cannot directly provide a clear explanation of the total, direct and indirect effects. When each variable has more than two categories and there are interactions among variables, it becomes more difficult to give a direct explanation of the effects. Although the indirect effect can be defined by subtracting the direct effect from the total effect, the interpretability of the effects is important (Hagenaars, 1998).

In this paper, we provide a method of path analysis of categorical variables. In section 2, path analysis is discussed in structural logistic regression models without interactive terms. We give definitions of the direct, indirect and total effects, and explain these effects in terms of log odds ratios, uncertain differences and an inner product of an explanatory vector and a response variable. In section 3, a study on food choice of alligators as a numerical example is reanalysed to illustrate the present approach. Section 4 extends the discussion to structural logistic regression models with interactive terms. In the final section, a summary and a discussion on the present approach are mentioned.

2. Effects of explanatory variables in structural logistic regression models without interactive terms

Let \( X_i \) \((i = 1, 2, \ldots, k)\) be categorical variables having categories \( \{1, 2, \ldots, I_i\} \). Assume that the structural relationship between \( X_k \) and \( \{X_i, i = 1, 2, \ldots, k-1\} \) in Fig. 1 is expressed by a logistic regression model without interactive terms. Let \( X_{pa(k)} = (X_1, X_2, \ldots, X_{k-1})' \); and let \( p(x_k | x_{pa(k)}) \) be the conditional probability of \( X_k = x_k \) given \( X_{pa(k)} = x_{pa(k)} \equiv (x_1, x_2, \ldots, x_{k-1})' \). Then, the logistic model is given as follows:

\[
p(x_k | x_{pa(k)}) = \frac{\exp \left( \alpha_{x_k} + \sum_{i=1}^{k-1} \beta_{ix_i,x_k} \right)}{\sum_{x_k=1}^{I_k} \exp \left( \alpha_{x_k} + \sum_{i=1}^{k-1} \beta_{ix_i,x_k} \right)},
\]

where \( \alpha_{x_k} \) are intercept parameters and \( \beta_{ix_i,x_k} \) are the effect parameters of \( X_i \). For model identification, suitable constraints are placed on the parameters, e.g.

\[
\alpha_1 = 0,
\beta_{i1j} = 0 \quad (j = 1, 2, \ldots, I_i; i = 1, 2, \ldots, k - 1),
\beta_{ij1} = 0 \quad (j = 1, 2, \ldots, I_i; i = 1, 2, \ldots, k - 1).
\]

This logistic model is denoted by \( \text{Logit}[X_1, X_2, \ldots, X_{k-1}] \), where the variables \( X_i \) in \([\ ]\) indicate the interactive terms of the highest orders with respect to the variables. In considering structural relationships among variables concerned, logistic regression models are referred to as structural logistic regression models in the present paper. In model (2.1), we introduce the following dummy variables:

\[
X_{ij} = \begin{cases} 
1 & (X_i = j) \\
0 & (X_i \neq j)
\end{cases} \quad (j = 1, 2, \ldots, I_i; i = 1, 2, \ldots, k).
\]

Random dummy vectors \( X_i = (X_{i1}, X_{i2}, \ldots, X_{iI_i})' \) (i = 1, 2, \ldots, k) and the corresponding categorical variables \( X_i \) are identified. Below, the random
dummy vectors are used for convenience of the discussion. Let $X_{pa(k)} = (X_1', X_2', \ldots, X_{k-1}')'$ and let $p(x_k | x_{pa(k)}) = P(X_k = x_k | X_{pa(k)} = x_{pa(k)})$. Model (2.1) is rewritten as

$$(2.2) \quad p(x_k | x_{pa(k)}) = \frac{\exp(x_k'\alpha + \sum_{i=1}^{k-1} x_k'B_i x_i)}{\sum_{x_k} \exp(x_k'\alpha + \sum_{i=1}^{k-1} x_k'B_i x_i)},$$

where

$$\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{I_k})'$$

and

$$B_i = \begin{pmatrix} \beta_{i11} & \beta_{i12} & \cdots & \beta_{i11} \\ \beta_{i21} & \beta_{i22} & \cdots & \beta_{i21} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{iI_k1} & \beta_{iI_k2} & \cdots & \beta_{iI_k1} \end{pmatrix} \quad (i = 1, 2, \ldots, k - 1).$$

The log odds ratio of $X_k = x_k$ over $X_k = x_k^*$ is given by

$$\log \frac{p(x_k | x_{pa(k)})}{p(x_k^* | x_{pa(k)})} - \log \frac{p(x_k^* | x_{pa(k)})}{p(x_k^* | x_{pa(k)})} = \sum_{i=1}^{k-1} \text{tr} B_i (x_i - x_i^*) (x_k - x_k^*)'.$$

This quantity implies an increase of the log odds at $X_{pa(k)} = x_{pa(k)}$ over baseline $X_{pa(k)} = x_{pa(k)}$. We formally set the above quantity as

$$(2.3) \quad \log OR(x_k, x_k^* | x_{pa(k)}, x_{pa(k)}^*) = \sum_{i=1}^{k-1} \text{tr} B_i (x_i - x_i^*) (x_k - x_k^*)'.$$

where

$$OR(x_k, x_k^* | x_{pa(k)}, x_{pa(k)}^*) = \frac{p(x_k | x_{pa(k)})p(x_k^* | x_{pa(k)})}{p(x_k^* | x_{pa(k)})p(x_k | x_{pa(k)})}.$$

**Remark 1.** Since

$$\log \frac{p(x_k | x_{pa(k)})}{p(x_k^* | x_{pa(k)})} = \log p(x | x_{pa(k)}) - \log p(x_k^* | x_{pa(k)})$$

$$= (- \log p(x_k^* | x_{pa(k)})) - (- \log p(x_k | x_{pa(k)}))$$

$$= (\text{uncertainty of } X_k = x_k^* \text{ given } X_{pa(k)} = x_{pa(k)})$$

$$- (\text{uncertainty of } X_k = x_k \text{ given } X_{pa(k)} = x_{pa(k)}),$$

this quantity can be interpreted as the amount of information that implies a decrease in uncertainty of $X_k = x_k$, compared with that of $X_k = x_k^*$. Hence, log odds ratio (2.3) means the change of uncertainty differences. □

**Remark 2.** The log odds ratio (2.3) is the inner product of the predictor vector $\sum_{i=1}^{k-1} B_i (x_i - x_i^*)$ and the response vector $x_k - x_k^*$. □
When we formally substitute the baselines $x^*_{pa(k)}$ and $x_*^k$ in (2.3) for the expectations of $X_{pa(k)}$ and $X_k$, i.e. $\mu_{pa(k)} = (\mu_1', \mu_2', \ldots, \mu_k-1')'$ and $\mu_k$ respectively, we have

$$\log OR(x_k, \mu_k; x_{pa(k)}, \mu_{pa(k)}) = \sum_{i=1}^{k-1} \text{tr} B_i(x_i - \mu_i)(x_k - \mu_k)'.$$

The above quantity can be viewed as the log odds ratio with respect to $(x_{pa(k)}, x_k)$, $(\mu_{pa(k)}, x_k)$, $(x_{pa(k)}, \mu_k)$, and $(\mu_{pa(k)}, \mu_k)$, and $\log OR(x_k; \mu_k; x_{pa(k)}, \mu_{pa(k)})$ is referred to as the log odds ratio based on mean baselines. First, the total effect of $X_{pa(k)} = x_{pa(k)}$ on $X_k = x_k$ is defined by

$$e_T(x_{pa(k)} \rightarrow x_k) = \log OR(x_k, \mu_k; x_{pa(k)}, \mu_{pa(k)})$$

$$= \sum_{i=1}^{k-1} \text{tr} B_i(x_i - \mu_i)(x_k - \mu_k)' .$$

Secondly, we define the direct effect of $X_i$ on $X_k$. In (2.4), if we formally set $x_{pa(k)} = (\mu_1', \mu_2', \ldots, \mu_{i-1}', x_i', x_{i+1}', \ldots, x_{k-1}')'$ and $x^*_{pa(k)} = (\mu_1', \mu_2', \ldots, \mu_{i-1}', \mu_i', x_{i+1}', \ldots, x_{k-1}')'$, then we have

$$\log OR(x_k, \mu_k; x_{pa(k)}, x^*_{pa(k)}) = \text{tr} B_i(x_i - \mu_i)(x_k - \mu_k)' .$$

This quantity can be regarded as the partial log odds ratio with respect to $X_i$ and $X_k$ given other explanatory variables, and is denoted by $\log OR(x_k, \mu_k; x_{pa(i)}, x_{i+1}, \ldots, x_{k-1})$. Then, the direct effect of $X_i = x_i$ on $X_k = x_k$ given $X_j = x_j$ ($j = i + 1, i + 2, \ldots, k - 1$) is defined by

$$e_d(x_i \rightarrow x_k | x_{i+1}, \ldots, x_{k-1}) = \log OR(x_k, \mu_k; x_i, \mu_i | \mu_{pa(i)}, x_{i+1}, \ldots, x_{k-1}) .$$

**Remark 3.** In the direct effect defined above, the direct effect of $X_{k-1} = x_{k-1}$ on $X_k = x_k$ is

$$e_d(x_{k-1} \rightarrow x_k) = \log OR(x_k, \mu_k; x_{k-1}, \mu_{k-1} | \mu_{pa(k-1)}) .$$

**Remark 4.** In the present case,

$$e_d(x_i \rightarrow x_k | x_{i+1}, \ldots, x_{k-1}) = \text{tr} B_i(x_i - \mu_i)(x_k - \mu_k)' ,$$

and the direct effect is independent of $X_j = x_j$ ($j = i + 1, i + 2, \ldots, k - 1$). Thus, we can write the direct effects as $e_d(x_i \rightarrow x_k)$ ($i = 1, 2, \ldots, k - 2$).

Thirdly, the total effect of $X_i$ on $X_k$ is discussed in model (2.2). The total effect of $(X_i, X_{i+1}, \ldots, X_{k-1}) = (x_i, x_{i+1}, \ldots, x_{k-1})$ on $X_k = x_k$ is defined by

$$e_T\{(x_i, x_{i+1}, \ldots, x_{k-1}) \rightarrow x_k\}$$

$$= \log OR(x_k, \mu_k; x_i, x_{i+1}, \ldots, x_{k-1}), (\mu_i, \mu_{i+1}, \ldots, \mu_{k-1}) | \mu_{pa(i)})$$

$$= \sum_{j=1}^{k-1} \text{tr} B_j(x_j - \mu_j)(x_k - \mu_k)' .$$
By replacing $\mathbf{x}_j$ ($j = i+1, i+2, \ldots, k-1$) in the above log odds ratio by the conditional expectation given $\mathbf{X}_i = \mathbf{x}_i$, i.e. $\mu_j(\mathbf{x}_i)$, the total effect of $\mathbf{X}_i = \mathbf{x}_i$ on $\mathbf{X}_k = \mathbf{x}_k$ is defined by

$$e_T(\mathbf{x}_i \rightarrow \mathbf{x}_k) = \log OR\{\mathbf{x}_k, \mu_k; (\mathbf{x}_i, \mu_{i+1}(\mathbf{x}_i), \ldots, \mu_{k-1}(\mathbf{x}_i))\},$$

$$\{\mu_i; \mu_{i+1}, \ldots, \mu_{k-1} \} \mid \mu_{pa(i)}\}$$

$$= \text{tr} B_i(\mathbf{x}_i - \mu_i)(\mathbf{x}_k - \mu_k)' + \sum_{j=i+1}^{k-1} \text{tr} B_j(\mu_j(\mathbf{x}_i) - \mu_j)(\mathbf{x}_k - \mu_k)'$$

$$(i = 1, 2, \ldots, k-2).$$

The first term in the above equation is the direct effect of $\mathbf{X}_i = \mathbf{x}_i$ on $\mathbf{X}_k = \mathbf{x}_k$, so the indirect effect is defined by the second term:

$$e_{ind}(\mathbf{x}_i \rightarrow \mathbf{x}_k) = \sum_{j=i+1}^{k-1} \text{tr} B_j(\mu_j(\mathbf{x}_i) - \mu_j)(\mathbf{x}_k - \mu_k)' = \log OR\{\mathbf{x}_k, \mu_k; (\mathbf{x}_i, \mu_{i+1}(\mathbf{x}_i), \ldots, \mu_{k-1}(\mathbf{x}_i))\},$$

$$\{\mathbf{x}_i, \mu_{i+1}, \ldots, \mu_{k-1} \} \mid \mu_{pa(i)}\}.$$

**Remark 5.** With respect to the effect of $\mathbf{X}_{k-1} = \mathbf{x}_{k-1}$ on $\mathbf{X}_k = \mathbf{x}_k$,

$$e_d(\mathbf{x}_{k-1} \rightarrow \mathbf{x}_k) = e_T(\mathbf{x}_{k-1} \rightarrow \mathbf{x}_k).$$

In the above consideration, the direct, indirect and total effects can be interpreted through log odds ratios, the inner product and the uncertainty difference. With respect to the indirect effects, we have the following theorem:

**Theorem 1.** If $\mathbf{X}_i$ and $\mathbf{X}_j$ are independent ($j = i+1, i+2, \ldots, k-1$), then

$$e_{ind}(\mathbf{x}_i - \mathbf{x}_k) = 0.$$

**Proof.** If $\mathbf{X}_i$ and $\mathbf{X}_j$ ($j = i+1, i+2, \ldots, k-1$) are independent, we get

$$\mu_j(\mathbf{x}_i) = \mu_j.$$

From (2.6) the theorem follows. $\square$

Lastly, the average effects are defined in order to summarize the effects defined above. The expectation of (2.5) is

$$E\{e_T(\mathbf{X}_i \rightarrow \mathbf{X}_k)\} = \text{tr} B_i \text{Cov}(\mathbf{X}_i, \mathbf{X}_k) + \sum_{j=i+1}^{k-1} \text{tr} B_j \text{Cov}(\mu_j(\mathbf{X}_i), \mu_k(\mathbf{X}_i)),$$

where

$$\text{Cov}(\mathbf{X}_i, \mathbf{X}_k) = E\{(\mathbf{x}_i - \mu_i)(\mathbf{x}_k - \mu_k)\}.$$
PATH ANALYSIS WITH LOGISTIC REGRESSION MODEL

Figure 2. Path Diagram of Lake, Size and Food

Table 1. Data of Primary Food Choice of Alligators by Lake and Size

<table>
<thead>
<tr>
<th>Lake</th>
<th>Size</th>
<th>Food</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hancock</td>
<td>small(≤2.3m)</td>
<td>Fish</td>
<td>Invertebrate</td>
<td>Reptile</td>
<td>Bird</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Hancock</td>
<td>large(&gt;2.3m)</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Oklawaha</td>
<td>small(≤2.3m)</td>
<td>5</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Oklawaha</td>
<td>large(&gt;2.3m)</td>
<td>13</td>
<td>8</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Trafford</td>
<td>small(≤2.3m)</td>
<td>5</td>
<td>11</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Trafford</td>
<td>large(&gt;2.3m)</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>George</td>
<td>small(≤2.3m)</td>
<td>16</td>
<td>19</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>George</td>
<td>large(&gt;2.3m)</td>
<td>17</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

\[ \text{Cov}(\mu_j(X_i), \mu_k(X_i)) = E\{(\mu_j(X_i) - \mu_j)(\mu_k(X_i) - \mu_k)\}' \]

This is the average total effect of $X_i$ on $X_k$, and the quantity is the inner product of random vectors $X_k$ and $\text{tr}(B_iX_i + \sum_{j=i+1}^{k-1} \text{tr}(B_j\mu_j(X_i)))$. This can be viewed as the covariance of them. The first term is the average direct effect of $X_i$ on $X_k$:

\[ E\{e_d(X_i \rightarrow X_k)\} = \text{tr}(B_i \text{Cov}(X_i, X_k)), \]

and the second one is the average indirect effect:

\[ E\{e_{ind}(X_i \rightarrow X_k)\} = \sum_{j=i+1}^{k-1} \text{tr}(B_j \text{Cov}\{\mu_j(X_i), \mu_k(X_i)\}). \]

3. Numerical example

Table 1 shows the data for an investigation of factors influencing the primary food choice of alligators (see Agresti, 1990, pp. 307–310). In this example, $X_1 = \text{Lake}$: lakes where alligators live; $X_2 = \text{Size}$: sizes of alligators; and $X_3 = \text{Food}$: primary food choice of alligators. The structural relationship among the variables are shown in Fig. 2. Firstly, the structural relationship between Food and (Lake, Size) is considered. The logistic regression model is assumed to be $\text{Logit}(\text{Lake, Size})$, and SPSS is employed for the parameter estimation (SPSS Advanced Statistics 7.5J, 1997). The estimated effect parameters $\beta$ in $\text{Logit}(\text{Lake, Size})$ are shown in Table 2, and the effects of factors, Lake and Size, are calculated by using the above method (Table 3). This table shows that the indirect
Table 2. Estimated Effect Parameters $\beta$ in Logit [Lake, Size]

<table>
<thead>
<tr>
<th>Food</th>
<th>Small</th>
<th>Fish</th>
<th>Invertebrate</th>
<th>Reptile</th>
<th>Bird</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale</td>
<td>Hancock</td>
<td>$-0.332$ (0.448)</td>
<td>$1.127$ (0.505)</td>
<td>$-0.683$ (0.651)</td>
<td>$-0.962$ (0.713)</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Oklawaha</td>
<td>$-0.826$ (0.558)</td>
<td>$-2.485$ (0.743)</td>
<td>$0.417$ (1.261)</td>
<td>$-0.131$ (0.892)</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Trafford</td>
<td>$-2.316$ (0.621)</td>
<td>$-0.394$ (0.626)</td>
<td>$1.419$ (1.189)</td>
<td>$-0.429$ (0.938)</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>George</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
</tbody>
</table>

*)The values in parentheses are asymptotic standard errors of the estimators.

*)The values 0.000 are the fixed values for the parameter estimation.

Table 3. Effects of Size and Lake on Food

<table>
<thead>
<tr>
<th>Size</th>
<th>Lake</th>
<th>Food</th>
<th>Hancock</th>
<th>Oklawaha</th>
<th>Trafford</th>
<th>George</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>Hancock</td>
<td>$-0.168$</td>
<td>$0.465$</td>
<td>$-0.320$</td>
<td>$-0.441$</td>
<td>$-0.024$</td>
</tr>
<tr>
<td></td>
<td>Oklawaha</td>
<td>$0.219$</td>
<td>$-0.607$</td>
<td>$0.148$</td>
<td>$0.576$</td>
<td>$0.031$</td>
</tr>
<tr>
<td></td>
<td>Trafford</td>
<td>$0.446$</td>
<td>$-1.273$</td>
<td>$0.127$</td>
<td>$0.846$</td>
<td>$0.696$</td>
</tr>
<tr>
<td></td>
<td>George</td>
<td>$-0.942$</td>
<td>$0.694$</td>
<td>$0.715$</td>
<td>$-1.130$</td>
<td>$-0.753$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-0.559$</td>
<td>$0.462$</td>
<td>$0.774$</td>
<td>$0.194$</td>
<td>$0.341$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$0.254$</td>
<td>$0.193$</td>
<td>$-1.301$</td>
<td>$-0.041$</td>
<td>$-0.322$</td>
</tr>
<tr>
<td>Large</td>
<td>Hancock</td>
<td>$0.055$</td>
<td>$0.153$</td>
<td>$-0.160$</td>
<td>$-0.145$</td>
<td>$-0.008$</td>
</tr>
<tr>
<td></td>
<td>Oklawaha</td>
<td>$0.058$</td>
<td>$0.110$</td>
<td>$-0.121$</td>
<td>$0.152$</td>
<td>$0.008$</td>
</tr>
<tr>
<td></td>
<td>Trafford</td>
<td>$0.044$</td>
<td>$0.084$</td>
<td>$-0.062$</td>
<td>$0.115$</td>
<td>$0.006$</td>
</tr>
<tr>
<td></td>
<td>George</td>
<td>$-0.033$</td>
<td>$0.091$</td>
<td>$-0.086$</td>
<td>$-0.086$</td>
<td>$-0.005$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$-0.322$</td>
<td>$-0.086$</td>
<td>$-0.086$</td>
<td>$-0.086$</td>
<td>$-0.005$</td>
</tr>
<tr>
<td>Average</td>
<td>Direct</td>
<td>$0.085$</td>
<td>$0.221$</td>
<td>$-0.015$</td>
<td>$0.206$</td>
<td></td>
</tr>
<tr>
<td>Effect</td>
<td>Indirect</td>
<td>$-0.008$</td>
<td>$0.008$</td>
<td>$0.006$</td>
<td>$-0.005$</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Estimated Effect Parameters $\beta$ in Logit [Lake]

<table>
<thead>
<tr>
<th>Lake</th>
<th>Hancock</th>
<th>Oklawaha</th>
<th>Trafford</th>
<th>George</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>$0.261$ (0.393)</td>
<td>$-0.942$ (0.390)</td>
<td>$-0.798$ (0.379)</td>
<td>0.000</td>
</tr>
<tr>
<td>Large</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
<td>$0.000$</td>
</tr>
</tbody>
</table>

*)The values in parentheses are asymptotic standard errors of the estimators.

*)The values 0.000 are the fixed values for the parameter estimation.

effects of Lake are much smaller than the direct effects. With respect to the direct effects of Lake, we can mention as follows. Alligators in Hancock primarily eat birds and others, alligators in Ocklawaha invertebrates and reptiles, alligators in Trafford invertebrates and reptiles, and alligators in George fish, e.g. the partial odds of $Food = Bird$ over the $Food$ mean at Lake = Hancock is $\exp(0.846)=2.330$ times higher than at
the \textit{Lake} mean, the partial odds ratio of \textit{Food} = \textit{Other} over the \textit{Food} mean at \textit{Lake} = \textit{Hancock} is \( \exp(0.696) = 2.006 \) times higher than at the \textit{Lake} mean, the partial odds of \textit{Food} = \textit{Invertebrate} over the \textit{Food} mean at \textit{Lake} = \textit{Oklawaha} is \( \exp(0.696) = 2.002 \) times higher than at \textit{Lake} mean, etc. From the \textit{Size} effects, small size alligators tend to eat invertebrates, i.e. the partial odds of \textit{Food} = \textit{Invertebrate} over the \textit{Food} mean at \textit{Size} = \textit{Small} is \( \exp(0.465) = 1.592 \) times higher than at the \textit{Size} mean. On the other hand, large size alligators have a tendency to eat reptiles and birds, e.g. the partial odds of \textit{Food} = \textit{Reptile} over the \textit{Food} mean at \textit{Size} = \textit{Large} is \( \exp(0.418) = 1.519 \) times higher than at the \textit{Size} mean; and that of \textit{Food} = \textit{Bird} over the \textit{Food} mean at \textit{Size} = \textit{Large} is \( \exp(0.576) = 1.779 \) times higher than at the \textit{Size} mean.

In the indirect effects of \textit{Lake}, the following indirect effects are comparatively large:

\[
\begin{align*}
    e_{\text{ind}}(\text{Lake} = \text{Hancock} \rightarrow \text{Food} = \text{Invertebrate}) &= 0.153, \\
    e_{\text{ind}}(\text{Lake} = \text{Oklawaha} \rightarrow \text{Food} = \text{Reptile}) &= 0.110, \\
    e_{\text{ind}}(\text{Lake} = \text{Oklawaha} \rightarrow \text{Food} = \text{Bird}) &= 0.152, \\
    e_{\text{ind}}(\text{Lake} = \text{Trafford} \rightarrow \text{Food} = \text{Reptile}) &= 0.084, \\
    e_{\text{ind}}(\text{Lake} = \text{Trafford} \rightarrow \text{Food} = \text{Bird}) &= 0.115, \\
    e_{\text{ind}}(\text{Lake} = \text{George} \rightarrow \text{Food} = \text{Invertebrate}) &= 0.091.
\end{align*}
\]

Considering the direct effects of \textit{Size} on \textit{Food} = \textit{Bird}, \textit{Invertebrate} and \textit{Reptile}, it may be concluded that these results concerning the above indirect effects come from the fact that more small size alligators live in Hancock and George than in other lakes, and that more large size alligators live in Oklawaha and Trafford than in other lakes. Table 3 shows that the greater part of the effect on \textit{Food} comes from factor \textit{Lake}.

Secondly, the effects of \textit{Lake} on \textit{Size} are considered. We get the estimates of the parameters in \textit{logit\{Lake\}} (Table 4), while the effects of \textit{Lake} on \textit{Size} are shown in Table 5. From this table, more small size alligators live in Hancock and George more than in Trafford and Oklawaha, i.e. the odds of \textit{Size} = \textit{small} over the \textit{Size} mean at \textit{Lake} = \textit{Hankock} is \( \exp(0.258) = 1.294 \) times higher than at the \textit{Lake} mean; and that at \textit{Lake} = \textit{George} is \( \exp(0.145) = 1.156 \) times higher than at the \textit{Lake} mean. On the other hand, more large size alligators live in Oklawaha and Trafford than in other lakes. The odds of \textit{Size} = \textit{large} over the \textit{Size} mean at \textit{Lake} = \textit{Ocklawaha} is \( \exp(0.344) = 1.411 \) times higher than at the \textit{Lake} mean, and that at \textit{Lake} = \textit{Trafford} is \( \exp(0.263) = 1.301 \) times higher than at the \textit{Lake} mean.

In this numerical example, the usefulness of the present method has been demonstrated.
4. Effects of explanatory variables in structural logistic regression models with interactive terms

A general discussion is very complicated. For simplification, we go on with the discussion by using the following example. In Fig. 3, \( \text{Logit}[X_1X_2, X_3] \) is assumed in the structural relationship between \( X_4 \) and \( \{X_1, X_2, X_3\} \). Let

\[
X_1 \otimes X_2 \equiv (X_{11}X_{21}, X_{11}X_{22}, \ldots, X_{11}X_{2I_2});
\]

and let \( p(x_4 | x_{pa(4)}) \) be the conditional probability of \( X_4 = x_4 \) given \( X_{pa(4)} = x_{pa(4)} \). Then, the logistic model is

\[
(4.1) \ p(x_4 | x_{pa(4)}) = \frac{\exp \left( x_4 \alpha + \sum_{i=1}^{3} x_4'B_i x_i + x_4'B_{12}(x_1 \otimes x_2) \right)}{\sum_{x_4} \exp \left( x_4 \alpha + \sum_{i=1}^{3} x_4'B_i x_i + x_4'B_{12}(x_1 \otimes x_2) \right)},
\]

where

\[
\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{I_4})',
\]

\[
B_i = \begin{pmatrix}
\beta_{i11} & \beta_{i12} & \cdots & \beta_{i1I_2} \\
\beta_{i21} & \beta_{i22} & \cdots & \beta_{i2I_2} \\
\cdots & \cdots & \cdots & \cdots \\
\beta_{i41} & \beta_{i42} & \cdots & \beta_{i4I_2}
\end{pmatrix}
\]

\((i = 1, 2, 3),\)

\[
B_{12} = \begin{pmatrix}
\beta_{11(11)} & \beta_{11(12)} & \cdots & \beta_{11(I_2)} \\
\beta_{12(11)} & \beta_{12(12)} & \cdots & \beta_{12(I_2)} \\
\cdots & \cdots & \cdots & \cdots \\
\beta_{14(11)} & \beta_{14(12)} & \cdots & \beta_{14(I_2)}
\end{pmatrix}.
\]

The previous discussion is directly applied to this case. In this case,

\[
\log OR(x_4, x_4'; x_{pa(4)}, x_{pa(4)}') = \sum_{i=1}^{3} \text{tr} \left( B_i (x_i - x_i') (x_i - x_i')' \right) + \text{tr} B_{12} (x_1 \otimes x_2 - x_1' \otimes x_2') (x_4 - x_4')'.
\]

The total effect of \( (X_1', X_2', X_3') = (x_1', x_2', x_3') \) on \( X_4 = x_4 \) is

\[
(4.2) \ e_T(x_{pa(4)} \rightarrow x_4) = \log OR(x_4; \mu_4; x_{pa(4)}, \mu_{pa(4)})
\]

\[
= \sum_{i=1}^{3} \text{tr} B_i (x_i - \mu_i) (x_4 - \mu_4)' + \text{tr} B_{12} (x_1 \otimes x_2 - \mu_1 \otimes \mu_2) (x_4 - \mu_4)'.
\]
The direct and indirect effects of $X_i = x_i$ ($i = 1, 2, 3$) on $X_4 = x_4$ are calculated as follows. The direct effect of $X_1 = x_1$ on $X_4 = x_4$ given $(X_2, X_3) = (x_2, x_3)$ is defined by

$$e_d(x_1 \rightarrow x_4 \mid x_2, x_3) = \log OR(x_4, \mu_4; x_1, \mu_1 \mid x_2, x_3) = \text{tr } B_1(x_1 - \mu_1)(x_4 - \mu_4)' + \text{tr } B_{12}(x_1 \otimes x_2 - \mu_1 \otimes x_2)(x_4 - \mu_4)' .$$

Similarly,

$$e_d(x_2 \rightarrow x_4 \mid x_3) = \log OR(x_4, \mu_4; x_2, \mu_2 \mid x_1, x_3) = \text{tr } B_2(x_2 - \mu_2)(x_4 - \mu_4)' + \text{tr } B_{12}(x_1 \otimes x_2 - x_1 \otimes \mu_2)(x_4 - \mu_4)' ,$$

and

$$e_d(x_3 \rightarrow x_4) = \log OR(x_4, \mu_4; \mu_3 \mid x_1, \mu_2) = \text{tr } B_3(x_3 - \mu_3)(x_4 - \mu_4)' .$$

With respect to the total effect (4.2), we have the following decomposition.

$$(4.3) \quad e_T(x_{pa(4)} \rightarrow x_4) = \log OR\{x_1, \mu_4; (x_1, x_2, x_3), (\mu_1, \mu_2, \mu_3)\} = \log OR\{x_1, \mu_4; (x_1, \mu_2, \mu_3), (\mu_1, \mu_2, \mu_3)\} + \log OR\{x_4, \mu_4; (x_1, x_2, x_3), (\mu_1, \mu_2, \mu_3)\} .$$

In this decomposition, the first term can be denoted by $e_d(x_1 \rightarrow x_4 \mid \mu_2, \mu_3)$, and the second term is $e_d\{x_2, x_3 \rightarrow x_4 \mid x_1\}$. By replacing $x_2$ and $x_3$ in (4.3) by the conditional expectations $\mu_2(x_1)$ and $\mu_3(x_1)$, the total effect of $X_1 = x_1$ on $X_4 = x_4$ is defined by

$$(4.4) \quad e_T(x_1 \rightarrow x_4) = \log OR\{x_1, \mu_4; (x_1, \mu_2, \mu_3), (\mu_1, \mu_2, \mu_3)\} + \log OR\{x_4, \mu_4; (x_1, \mu_2, (x_1, \mu_3(x_1)), (\mu_1, \mu_2, \mu_3)\} .$$

The first and second terms in (4.4) can be formally denoted by $e_d(x_1 \rightarrow x_4 \mid \mu_2, \mu_3)$ and $e_d\{(\mu_2(x_1), \mu_3(x_1)) \rightarrow x_4 \mid x_1\}$, respectively. From this decomposition, we can set

$$e_d(x_1 \rightarrow x_4) \equiv e_d\{x_1 \rightarrow x_4 \mid (\mu_2, \mu_3)\}$$

$$= \text{tr } B_1(x_1 - \mu_1)(x_4 - \mu_4)' + \text{tr } B_{12}(x_1 \otimes \mu_2 - \mu_1 \otimes \mu_2)(x_4 - \mu_4)' ,$$

$$e_{ind}(x_1 \rightarrow x_4) \equiv e_d\{(\mu_2(x_1), \mu_3(x_1)) \rightarrow x_4 \mid x_1\}$$

$$= \sum_{i=2}^{3} \text{tr } B_i(\mu_i(x_1) - \mu_i)(x_4 - \mu_4)' + \text{tr } B_{12}(x_1 \otimes \mu_2(x_1) - x_1 \otimes \mu_2)(x_4 - \mu_4)' .$$

In this sense, the total effect is decomposed into the direct and the indirect effects, and the effects can be interpreted as log odds ratios. Similarly, the total effect of $X_2 = x_2$ on $X_1 = x_1$ is defined by

$$e_T(x_2 \rightarrow x_4) = \log OR\{x_4, \mu_4; (x_2, \mu_3(x_2)), (\mu_2, \mu_3) \mid x_1\} .$$

Since

$$\log OR\{x_4, \mu_4; (x_2, \mu_3(x_2)), (\mu_2, \mu_3) \mid x_1\} = \log OR\{x_4, \mu_4; (x_2, \mu_3), (\mu_2, \mu_3) \mid x_1\} + \log OR\{x_4, \mu_4; (x_2, \mu_3(x_2)), (\mu_2, \mu_3) \mid x_1\} ,$$
we set as follows:

\[ e_d(x_2 \rightarrow x_4) = \log OR\{x_4, \mu_4; (x_2, \mu_3), (\mu_2, \mu_3) \mid \mu_1\} \]
\[ = \text{tr} B_2(x_2 - \mu_2)(x_4 - \mu_4)' + \text{tr} B_{12}(\mu_1 \otimes x_2 - \mu_1 \otimes \mu_2)(x_4 - \mu_4)', \]
\[ e_{\text{ind}}(x_2 \rightarrow x_4) = \log OR\{x_1, \mu_4; (x_2, \mu_3(x_2)), (x_2, \mu_3) \mid \mu_1\} \]
\[ = \text{tr} B_3(\mu_3(x_2) - \mu_3)(x_4 - \mu_4)' . \]

With respect to the effects of \( x_3 \) on \( x_4 \),

\[ e_T(x_3 \rightarrow x_4) = e_d(x_3 \rightarrow x_4) = \log OR\{x_4, \mu_4; x_3, \mu_3 \mid \mu_1, \mu_2\}. \]

In a general case shown in Fig. 1, we calculate the total effect of \((X_i, X_{i+1}, \ldots, X_{k-1}) = (x_i, x_{i+1}, \ldots, x_{k-1})\) on \( X_k = x_k \) by

\[ e_T\{(x_i, x_{i+1}, \ldots, x_{k-1}) \rightarrow x_k\} \]
\[ = \log OR\{x_k, \mu_k; (x_i, x_{i+1}, \ldots, x_{k-1}), (\mu_i, \mu_{i+1}, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\}. \]

With respect to the above effect we have the following theorem.

**Theorem 2.** In a general logistic regression model,

\[ (4.5) \log OR\{x_k, \mu_k; (x_i, x_{i+1}, \ldots, x_{k-1}), (\mu_i, \mu_{i+1}, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\} \]
\[ = \log OR\{x_k, \mu_k; (x_i, \mu_{i+1}, \ldots, \mu_{k-1}), (\mu_1, \mu_2, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\} \]
\[ + \log OR\{x_k, \mu_k; (x_i, x_{i+1}, \ldots, x_{k-1}), (x_i, \mu_{i+1}, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\} . \]

**Proof.** In a logistic regression model, the partial log odds ratio \((4.5)\) can be expressed by

\[ \log OR\{x_k, x_k^*; (x_i, x_{i+1}, \ldots, x_{k-1}), (x_i^*, x_{i+1}^*, \ldots, x_{k-1}^*) \mid \mu_{pa(i)}\} \]
\[ = \{g(\mu_{pa(i)}; x_i, x_{i+1}, \ldots, x_{k-1}) \]
\[ - g(\mu_{pa(i)}; x_i^*, x_{i+1}^*, \ldots, x_{k-1}^*)\}(x_k - x_k)', \]

where \( g(x_1, x_2, \ldots, x_{k-1}) \) is an appropriate function. From this, the theorem follows. \( \Box \)

In general cases, from the decomposition \((4.5)\) we can calculate the effects of \( x_i \) on \( x_k \) by

\[ (4.6a) \quad e_T(x_i \rightarrow x_k) = \log OR\{x_k, \mu_k; (x_i, \mu_{i+1}(x_i), \ldots, \mu_{k-1}(x_i)), \]
\[ \quad (\mu_1, \mu_{i+1}, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\}; \]
\[ (4.6b) \quad e_d(x_i \rightarrow x_k) = \log OR\{x_k, \mu_k; (x_i, \mu_i + 1, \ldots, \mu_{k-1}), \]
\[ \quad (\mu_1, \mu_2, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\}; \]
\[ (4.6c) \quad e_{\text{ind}}(x_i \rightarrow x_k) = \log OR\{x_k, \mu_k; (x_i, \mu_{i+1}(x_i), \ldots, \mu_{k-1}(x_i)), \]
\[ \quad (x_i, \mu_{i+1}, \ldots, \mu_{k-1}) \mid \mu_{pa(i)}\}. \]

In \((4.6a-c)\), we have the following decomposition:

\[ e_T(x_i \rightarrow x_k) = e_d(x_i \rightarrow x_k) + e_{\text{ind}}(x_i \rightarrow x_k). \]
In \((4.6c)\) if \(X_i\) and \(\{X_{i+1}, X_{i+2}, \ldots, X_{k-1}\}\) are independent,

\[
\mu_j(x_i) = \mu_j (j = i + 1, i + 2, \ldots, k - 1).
\]

From this,

\[
e_{\text{ind}}(x_i \rightarrow x_k) = 0.
\]

Hence, theorem 1 holds true in general cases. The advantage of the present approach is that the effects defined in \((4.6a-c)\) can be interpreted as log odds ratios.

5. Conclusion

In path analysis, the effects of explanatory variables on a response variable can be discussed in theory through a comparison of the conditional distributions of the response variable given explanatory ones, however a question arises as to how the total, direct and indirect effects are measured. Even if the effects are defined, a further problem in path analysis is how to interpret the total, direct and indirect effects. In path analysis of categorical variables the problems are more complicated than in continuous variables, because of multiple categories and interactive terms. This paper has provided a method of path analysis of categorical variables in fully recursive causal systems. The total, direct and indirect effects have been defined, and we have given a method of calculation of the effects. The effects defined in this paper can be interpreted in terms of (i) log odds ratios; (ii) changes of uncertainty of a response variable; and (iii) the inner products of explanatory and response variables. This is an advantage of the present method. It is important to extend the present approach to a method for treating more general causal systems of categorical variables; however it needs a further discussion following the present study.

Acknowledgements

The authors would like to thank the referee and the editor for their useful comments and suggestions for improving the first version of this paper. This research was supported by Grant-in-aid for Scientific Research 11680325, Ministry of Education, Culture, Sports, Science and Technology of Japan.

References


